ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Deduplication of Encoded Big Data in Cloud Storage

Pallavi G. Bangale, Prof Rekha Kulkarni

M.E. Student, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India.

Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India.

ABSTRACT: Cloud Computing makes a deeper impact on the world as it introduces the new paradigm for computing. Using a shared pool of the memories; users can expand or shrink the resources or memory according to the requirement. With Great power comes the great disadvantages also. As there is no control of Cloud Service Provider(CSP) to the data uploaded, users can upload same data on the Cloud causing the wastage of the memory. This problem of duplication of files can be reduced by the deduplication of files Deduplication occurs when the content of files such as identifier, token is known. If data is encrypted then the deduplication gets complicated. The paper represents the technique to deduplicate the files by using the proxy re-encryption technique. This is based on the encrypting data with another key. This approach contains only two participants including Cloud Service provider and the Clients which are users who try to upload and access files over the Cloud.

KEYWORDS: Big data, Cloud Computing, Data Deduplication

I. INTRODUCTION

Mobile Ad loud Computing is the distributed technology used for the ubiquitous from many machines located remotely. This technology enables to use the resources of the remote machines as their own and support the high performance computing by performing the applications on those machines, The backbone of the Cloud computing is the virtualization in which the virtual version of the resources are created for the computation. One type of the virtualization is the storage virtualization in which multiple physical storage systems from a single logical storage. This virtual storage has memory to store the big data and user unable to tell the difference between the physical and logical storage system, So as the Cloud Service Provider has to maintain the logical pool of such a large memory, he efficiency and performance of storage should be superior. As the data storage increases, the data management becomes complicated. The major reason behind the complex data management is the duplicate files reside on the Cloud. These files consumes the network, storage resources and also causes the energy wastage. As Cloud computing gains more popularity, multiple users have uploaded their data and that raise the problem of uploading the same data in the form of contents. Duplicate files cause deduction in the performance of the Cloud storage. As there are several unnecessary files over the Cloud, Cloud storage manager requires very efficient types of indexing to access these files and also maintaining such kind of data requires the costliest in terms of the economy and also in terms of energy. So to reduce the storage load over the cloud server, duplicate files have to be removed. Next problem occurred in removing the duplicate files, is the encryption of the files. Encryption of the file causes the conversion of the files into the cipher text. Different encryption methods causes the different cipher text of the same file. So if the same file is uploaded using the different encryption methods, the deduplication causes.

Deduplication of the data is one of the technique to reduce the cost, and storage demand for the Cloud application. It avoids the uploading of the duplicate files. Deduplication stores only one copy of the redundant data and provides links to that single copy irrespective of the number of clients requesting that file. Deduplication divides the file in chunks and then each chunk is compared with the data chunk in the Cloud storage to avoid the double uploading of the data. As the performance increases, more number of customers attract towards the technology, and security issues needs to be concerned then. As the unauthorised access to the confidential information and data is one of the most important concerns, Cloud Service Provider should secure the Cloud storage from malicious attacks performed by the eavesdroppers and third parties.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

In Practice the security to the data is provided by the cryptosystem. In this, the data is encrypted by using the keys by data sender and then it is decrypted by data receiver using the same or different keys depending on the type of encryption. Now as the data gets changes, then it is difficult to compare the data with the another data. Deduplication works only when the chunks of the data can be compared. So encrypted data produce new challenges for data deduplication. Also many attempts shows that deduplication of encrypted data can be achieved but they have to compromise with the security of the data. The additional servers are added to perform the deduplication of the data which further leads to the increased complexity of the structure. Cloud Service Provider also takes time to serve the request to the client because of the additional servers.

The paper proposes the system that works on the deduplication on the encrypted data without using any additional servers. The technique of the proxy re-encryption is going to be used in the system in case of the duplicate files. System would also take the help of the ownership challenge to identify the original data ownership of the uploaded file, Using the proxy re-encryption and the data ownership challenge, system can perform the data deduplication. This paper proposes the methods of deduplication and the proposed system architecture and analysis of the system. We are also using the Amazon AWS Cloud computing platform as the Cloud server and JSP as the programming environment for our work. The rest of the paper is organized as follows. Section II gives the brief overview of related work. Section III introduces the system. Section IV introduces the architecture and proposed work. Section V gives the result. Finally, a conclusion is presented in last section.

II. RELATED WORK

In Cloud storage service is provided by many organizations and institutes such as Google, Amazon, Dropbox which performs the data deduplication on the users data and then unnecessary upload is avoided. Google Drive uses the data ownership to identify the owner so that owner should have the additional authority to access the cloud storage. Whenever any modification is done it is only the file owner and all the other clients are then notified about the change. And also the old files are stored in the revision history of the files, with the default version is the updated one. Even Dropbox provides the data storage which provides the data deduplication. Experiments prove that uploading one. Experiments prove that uploading the same file again takes less time to upload, as at was never uploaded, just the ownership of the file is known to the clients so that the data can read, written by the permission of the owner. Reduction of time is the result of the less bandwidth requirements for the cloud access. Also the Amazon AWS provides the deduplication with the help of the StorReduce, which deals with petabytes of unstructured data for deduplication [5]. The removal of the unwanted files on Microsoft Azure is done by the Disk Deduplication in which Windows Server performs the data deduplication on virtual hard disks that are attached to virtual machines as backup[4].

Combining the deduplication and encryption is the hot topic in the research these days. Many researchers have developed methods to achieve the deduplication of encrypted data but they lack the security issue that needs to be considered. Such that in paper [1], author proposed a system to deduplicate which uses the additional server authorized party which is trusted and used by the cloud clients for re-encryption. But additional of extra server costs much and also complexity is increased. Another work [6] proposed the system that use two cloud servers, one for the storage and one for fingerprint which is the hash value of the data. Again another additional resources unnecessarily leads to the deduplication on block-level and they used extra component for the key management among the other clients and the data owner.

Also the system [8] developed by Zheng Yan used attribute based encryption to deduplicate the encrypted data. But attribute-based-encryption also suffers challenges like key coordination, key revocation. Also they lack non-efficiency and non-existence of attribute revocation method. Also the another system [9] propose the deduplication and proxy reencryption and proof of ownership to deduplicate the file.

In order to reduce workloads, another system is proposed which uses Index Name Server(INS) [10] to manage not only the data deduplication but also the compression of file, IP information. Another work [11], TIN-Yu-Wu proposed same Index Name Server which integrates data deduplication with the facility of the automation in the reduction of the numbers. In the given INS system we cannot distinguish between the different file formats. So we have the same bandwidth for each type of the file including the text, audio, video files. This will cause unequal balancing to the same bandwidths of the different file types. Suppose any request to the server contains the access to the text but slow to the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

video files. Then the same bandwidth will be provided to these two files. This will cause the fast access to text files. So current INS server cannot deduplicate the encrypted data. C. Fan and S. Y. Huang proposed a data deduplication [13] on hybrid cloud which supports deduplication on plain text and ciphertext. But this mechanism cannot support encrypted data deduplication very well. It assumes that CSP knows the encryption key, but that's not the case always. CSP cannot be fully trusted by the data holders and owners. Many developers use Elliptic Curve Cryptography (ECC),Proxy Re-encryption(PRE), or any asymmetric curve cryptography. But in this system we used the Advanced Encryption Standard (AES) which is also symmetric encryption algorithm. It has several advantages over other encryption methods like it is more secure and it also supports larger key sizes. Also it is more faster.

III. PROPOSED ALGORITHM

Proposed system follows client-server architecture. Client consist of the users which can upload the data which consist of the files which contains the text files. Server consist of the CSP which accepts the request from the client and then the server runs the deduplication algorithm and then server sends the response to the client. Server accepts the request which consists of the encrypted data, and when the deduplication algorithm occurs, the data is stored in the cloud storage. Then CSP sends response to the client in the form of notification. This section discusses the architecture of the proposed system. We propose a solution that deduplicates the data irrespective of its encryption technique. The architecture of the system is shown in figure 1. There are two types of the users in our system, Data owner and user. Data owner actually uploads the file and has ownership of that file. The user is one who tries to upload the file and has an ownership of that file. The user is one who tries to upload the same file but because of deduplication it will not get uploaded. So we have to share the key with the owner so that afterward user can also download the file. This system has client and server side and controller is the bridge between them. Controller accepts the request from the client and user and then further sends to re-encryption to check the duplication. This system contains the main aspect of identifying the duplicate file. It will identify the file by using fingerprints which are the unique IDs assigned to each data chunk. Data chunks are blocks of the data also known as the tokens. So as the tokens are uploaded, they are compared with the already uploaded tokens. If the token match is positive, it means that data duplication occurred.

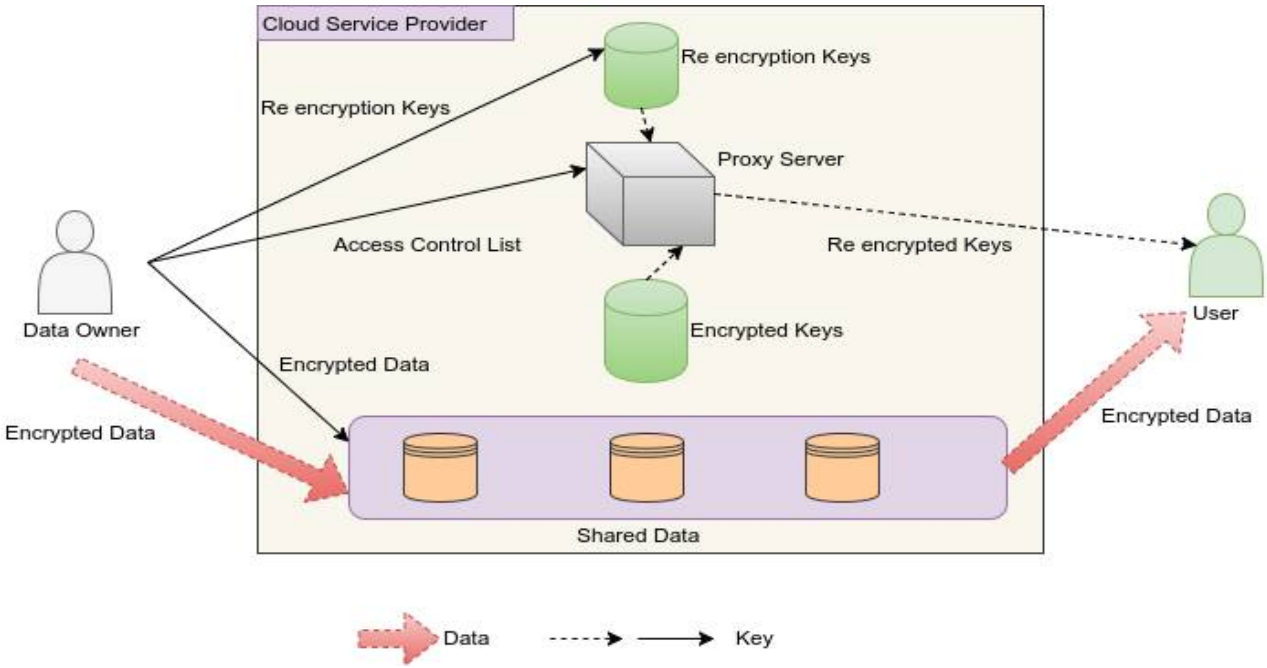

Fig 1. Architecture of the system

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

When deduplication occurs, then Cloud Service Provider challenges the data ownership challenge, which checks the eligibility of the data holder, then it applies encryption again with the new key so that only the eligible data holder can have that key and only that data holder can decrypt that file. This technique of encrypting the data again with new key is called as the re-encryption and all the authorized data holders have the re-encryption key. When the data owner tries to delete the files, then CSP does not delete the file actually, it just makes the data owner unauthorized so that he no longer access the file. If that file has no duplication record which means that if that file not tried to upload by any other user, then that file will be deleted from Cloud Storage. In case, The data owner updates the original file such that the encryption key for that file is changed, then all the other data holders are update about it by the Cloud Service Provider. As we are using the AES encryption technique for the system it is more secure that the other encryption techniques.

Additional assumptions we include are: Users provide the encrypted hash codes of the data for ownership verification. The data owner has highest priority. Users should provide a valid certificate in order to request special treatment. Users and CSP communicate through a secure channel with each other CSP can authenticate its users in the process of Cloud data storage.

IV. RESULTS AND DISCUSSION

The proposed system consist of the main logon page which consist of the login for two types of users, the end user and the CSP. Login ID and password is provided for the each user and also for the CSP. When user logs in the system, user can upload the file by clicking on the 'upload' button and also search the file on clicking 'search' button. User also download the file by clicking on 'Document Request' button and then it can decrypt document by using the 'Decrypting Document' button. In the upload view button, user can see the details of the uploaded data. User can also log out application on clicking 'Quit' button. CSP home page consist of the 'Cloud user details' in which all the details are shown. Then the 'Cloud Storage Document Details' link consist of the list of all documents uploaded by the user and 'Quit' button is provided for log out of CSP account. Proposed system can be compared in terms of the time required foe server to respond. Less time is efficient for big data and encrypted data. AES is very efficient in terms of response time from server to client and vice versa. The figure 2 shows the access frequency of data files over the storage in a monthly duration. The graph shows that over a period of month, a particular file is access how many times. The X-axis shows the documentId of the document and Y-axis shows the number of users who downloaded the file or tried to access the file. As we can see, the document 4 is tried to access by 4 users over a period of month. Which also means that documentId 4 is tried to uploaded by 4 different users. This information provides us the rate at which particular duplicate document is tried to upload by different users. This provides the duplication of any document. This access rate can be used to decide the most accessed document over a given period which can be further useful to Cloud Service Provider to give the solution to change the bandwidth for the most accessed document so that the user can access or download it in very less time. This particular access frequency is useful in many social networking sites for deciding the trend of the particular period.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

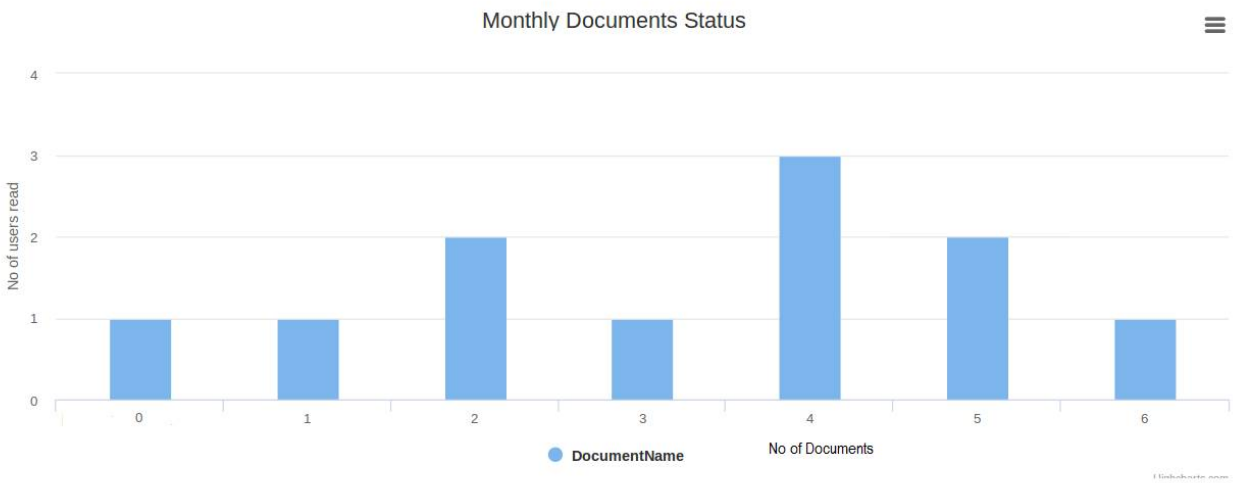

figure 2 : Access Frequency of Data Files

The performance of the system in terms of encryption time for number of 16-bytes block is shown in figure 3. For that we compared the 4 encryption technique and their performance for the different number of blocks. Among them, Triple DES takes most time when number of blocks increases. Proposed approach uses the AES of 128 bits which takes less time compared with all other approach. As the number of blocks increases the encryption speed varies from 0.5 milliseconds to 10 milliseconds. But other encryption techniques like DES and triple DES increases linearly up to 110 milliseconds. So using AES encryption reduces much time for the encryption when the big data is concerned.

V. CONCLUSION AND FUTURE WORK

The Huge cloud size and large number of end users cause the duplication of the files and so it is hard to manage the storage and it degrades the performance. Therefore deduplication of data needs to be done. Managing encrypted data is also important in practice for achieving a successful cloud storage service. This system works on a data ownership and proxy re-encryption method to deduplicate the encrypted file. Using only cloud service provider to achieve deduplication we propose a method to avoid the duplication of data encrypted using the AES encryption method. This work is to implement the deduplication of files over Amazon AWS Cloud computing platform. Over that deploying virtual machines and then deploying a datastore to store the files will be in the experimental work. By taking the AES 128 bit encryption, time taken to encrypt and deduplicate the data become less. Using PAKE protocol, the deduplication can be achieved that enables two parties to privately compare their data and share the encryption key.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

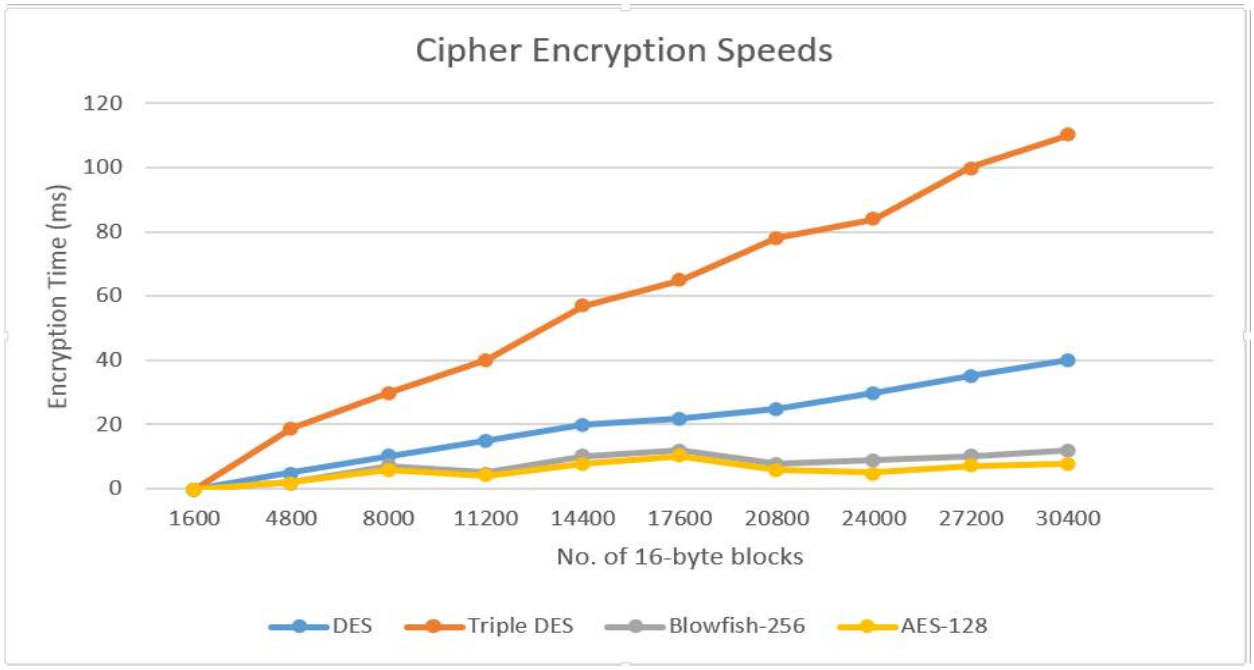

figure 3 : Cipher Encryption Speed

REFERENCES

1. Zheng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, Robert H. Deng, "Deduplication on Encrypted Big Data in Cloud ," IEEE Transaction on Big Data, April-June 2016,pp. 138-150.
2. Hur, Junbeom, et al. "Secure data deduplication with dynamic ownership management in cloud storage." IEEE Transaction on Knowledge and Data engineering 28.11 (2016): 3113-3125.
3. http://opendedup.org/, "Opendedup".(2016).[Online] Available: http://opendedup.org/odd.
4. docs.microsoft.com, "Azure backup", 2016 [Online]. Available: https://docs.microsoft.com/en-us/azure/backup/backup-introduction-toazure-backup.
5. aws.amazon.com, "Cloud Deduplication, On -Demand: StorReduce , an APN Technology Partner," 2016. [Online]. Available: https://aws.amazon.com/blogs/apn/Cloud-deduplication-on-demandstrirreduce-an-apn-technology-parthner.
6. Zhaocong wen,Jinman Luo, Huajun Chen, Jiaxiao Meng, Xuan Li, Jin Lim "A Verifiable Data Deduplication Scheme in Cloud Computing" Intelligent Networking and Collaborative System (INCoS), 2014 International Conference (2015): 85-90.
7. Pasqual Ouzio, Refik Malva, Melek Onen, Sergio Loureiro, "Cloudedup: SEcure Deduplication with Encrypted Data for Cloud Storage," cloud computong Technology and science (CloudCom), 2013 IEEE 5th International Conference(2013):363-370.
8. Zheng Yan, Mingjun Wang, Yuxiang Li, Athanasios V. Vasilakos, "Encrypted Data Management with Deduplication in Cloud Computing," IEEE Cloud computing (2016):28-35.
9. Xuexue Jin, Lingbo Wei, Mengke Yu, Nenghai Yu, Jinyuan Sun, "Anonymous deduplication of rncrypted data with proof of ownership in cloud storage," Communication in China(ICCC),2013 IEEE/CIC International Conference (2013)224-229.
10. Wu, Tin-Yu, Jeng-Shyang Pan, and Chia-Fan Lin. "Improving accessing efficiency of cloud storage using e-duplication and feedback schemes" IEEE System Journal 8.1 (2014): 208-218.
11. Fu, Kevin, Seny Kamara, and Tadayoshi Kohno. "Key regression: Enabling efficient key distribution for secure distributed storage," Computer Science Department Faculty Publication Series (2006): 149.
12. Li, Jingwei, et al. "REkeying for encrypted deduplication storage." Dependable System and Networks (DSN), 2016 46th Annual IEEE/IFIP International Conference on. IEEE, (2016): 618-629.
13. Liu, Jian, N. Asokan and Benny Pinkas. " secure deduplication of encrypted data without additional independent servers," Proceeding of the 22nd ACM SIGSAC COnference on Computer and Communications security (2015):874-885.
14. Peterson, Zachary NJ, et al. "Secure Deletion for a versioning File System." FAST vol. 5. No. 2005, 2005.
15. Storer, Mark W. et al. "SEcure data deduplication." Proceeding of the 4th ACM international workshop on storage security and survivability.(2008): 1-10.